DESIGNING A PRIVACY-PRESERVING GEOVISUALIZATION OF CITIZEN-COLLECTED ENVIRONMENTAL DATA

By

ADDISON ARMSTRONG KAUFMANN

_____

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Computer Science

THE UNIVERSITY OF ARIZONA

M A Y 2 0 1 8

Approved by:

_____

Dr. Kate Isaacs
Department of Computer Science

Abstract

This thesis project is a web-based map of environmental data from a citizen science study. The data includes soil and water contamination values from collection sites across four Arizona counties. The main priority of the visualization was to protect the privacy of the initial study participants, while still providing an engaging and accurate mapping. This objective was addressed by using census block groupings to obscure the exact location of the data collection sites, and coloring those blocks with the average values. Other design considerations included how to handle the color scale across the different maps, how to communicate the magnitude of the data, and how to represent the distribution of data within a grouping. Each of these design considerations was explored in detail before completing the final design. The next steps for this project include replication with a similar data set, and final deployment on the web.

Introduction

This thesis explores how to design an engaging visualization of environmental data from a citizen science study. The goal was to design and build a web map that communicates the data and inspires interest in the study, while protecting the privacy of the study participants. The data behind this thesis comes from Gardenroots: A Citizen Science Project, conducted by the Integrated Environmental Science and Health Risk Laboratory (IEHSRL) at University of Arizona. Gardenroots was a citizen science study, meaning that members of the general public collected the data, as part of a collaboration with professional scientists at this laboratory. Citizen science is vitally important, especially in the area of environmental health. Environmental health is the science and practice of protecting public health by identifying and studying environmental hazards, and limiting exposure to such hazards in air, water, soil, plants and food. In this field, involving citizens in scientific practice helps inform those citizens about the potential health risks they face in their daily lives. In Arizona, these health risks are most prevalent in rural communities as a result of mining waste pollution. Gardenroots focuses specifically on these communities, and the health of their soil and water.

This visualization is an extension of the Gardenroots study. The main goal was to build an interactive and interesting web map that would help communicate the data back to those initial participants, and to the general public. We hope that this visualization engages viewers with citizen science and environmental health, while encouraging them to participate in similar studies in the future.

This project had some challenges and risks associated with designing the visualization. Because this study collected sensitive data from individuals, privacy and security for those participants was a crucial priority. The other main priority was to accurately represent the science behind this data. We did not want to alarm users by overstating the risks associated with soil contamination. To ensure that both of these issues were addressed appropriately, regular meetings were held during each step of the design with Dr. Ramirez-Andreotta, the IEHSRL principal investigator. This thesis was also advised by Dr. Isaacs, who specializes in data visualization in the Department of Computer Science. Dr. Isaacs provided guidance on the design and technical decisions of the project.

Background

Gardenroots is a study that engages community members through citizen science about the health of their soil, water, and plants. The IEHSRL in University of Arizona College of Agriculture and Life Sciences conducted this study in partnership with citizens in Apache, Cochise, Yavapai, and Greenlee counties. Within these counties, Gardenroots focused on rural communities, which often rely on private wells that do not undergo the same rigorous contamination testing as do public wells in larger cities. This laboratory trained community members throughout these counties to collect soil and water samples from their homes and gardens. The laboratory then analyzed these samples for concentrations of different minerals.

The minerals tested include beryllium, sodium, magnesium, aluminum, potassium, calcium, vanadium, chromium, manganese, iron, cobalt, nickel, copper, zinc, arsenic, selenium, molybdenum, silver, cadmium, tin, antimony, barium, and lead. Among these minerals, some are considered potentially hazardous at certain levels set by the Environmental Protection Agency (EPA). For soil samples, these hazard levels are

called “residential screening levels” and are based on research linking long-term exposure to health problems. The table below shows the residential screening levels of each potentially hazardous substance tested in this study.

Element	Residential Screening Level (ug/mg)
Beryllium	1,500
Aluminum	77,000
Vanadium	78
Chromium	30
Cobalt	900
Nickel	1,600
Copper	3,100
Zinc	23,000
Arsenic	10
Selenium	390
Molybdenum	390
Silver	390
Cadmium	39
Tin	47,000
Barium	15,000
Lead	400

Fig 1. A table showing the EPA residential screening level for each element tested in the Gardenroots study

After the analysis was completed, the study focused efforts on how to communicate these results back to the communities, and to a broader audience. First, the IEHSRL team designed booklets that were sent to each individual participant. The booklets showed each participants data compared with the residential screening level values, and compared with the other participants data as a whole. These booklets used common visualizations like bar charts, and were designed to make this data as accessible and easy-to-understand as possible. Additionally, IEHSRL published reports in scientific journals (detailed in the related work section). These avenues of communication are valuable, but their scope is limited. With such a focus on community involvement, this laboratory wanted to explore different avenues of communication that would display this study’s findings to a broader audience. The purpose of this thesis is to make the data accessible and engaging to a large audience, while still protecting the privacy of the study participants. I designed and built a

web-based map visualization that gives an overview of the study's findings and encourages users to explore the data in a new way.

Design

The most important issue when designing this visualization was how to protect the privacy of the study participants. This study included data on lead and arsenic contamination, which can be sensitive information. For example, if there is publicly available information that a certain home has high arsenic concentration in the water, it may lower the home value or affect the owner's ability to sell. We did not want to discourage future participants by publishing this sensitive information and linking it to these participants exact location. Thus, the main design decisions were about how to obscure the location of the participants homes and gardens while still making the visualization engaging and communicative. The design was validated after each iteration by meeting with the thesis advisors, Dr. Ramirez-Andreotta and Dr. Isaacs, and incorporating their feedback.

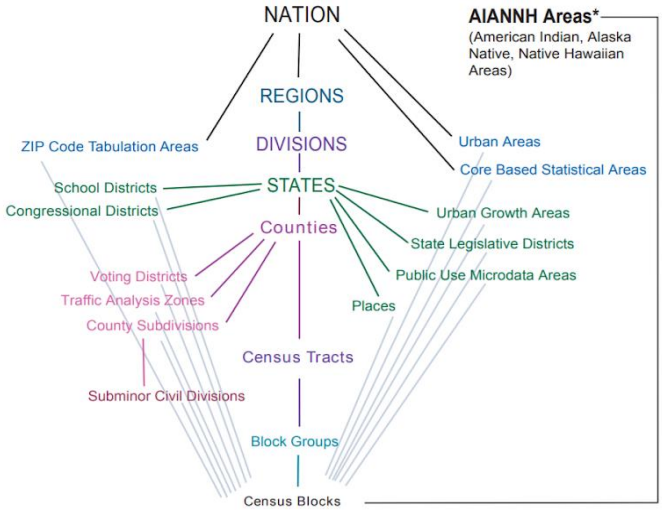

Fig 2. A graphic published by the US Census Bureau detailing the divisions used to collect census data.

First Iteration

The first step of design was obtaining a geographical unit large enough that viewers could not identify exact homes. The Census Bureau divides the United States into very small regions which they use to collect population and housing information. The smallest subdivision in the Census Bureau's system is called a census block. Each time a census is conducted, the divisions change slightly to reflect dynamic population and housing in each area. Blocks are usually bounded by streets or waterways. Census blocks can be as small as one city block, and as large as hundreds of acres in rural areas. Similarly, they can range in population from zero inhabitants to several hundred inhabitants. In the 2010 census, the Census Bureau identified more than 11 million census blocks across the United States, about 4 million of which were completely uninhabited.

All of the census blocks that contained one or more participant homes were collected. Overall, there were 398 samples taken, which were distributed across 162 census blocks. However, because Gardenroots focused heavily on rural areas, many of these census blocks contained fewer than ten household units. Figure 3 shows the distribution of number of households within the census blocks.

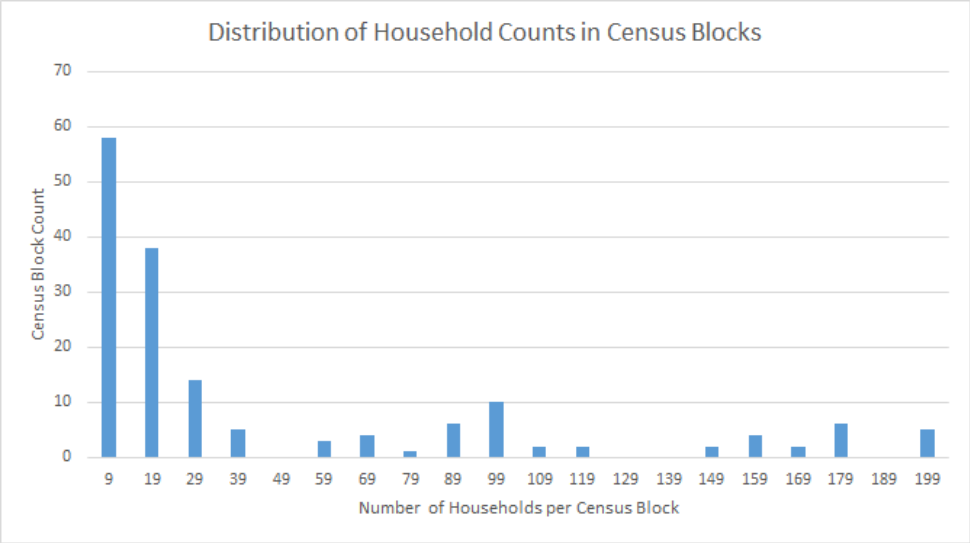

Fig 3. Distribution of number of households within each census block.
There were 58 census blocks with fewer than 10 households.

In those cases, a census block was insufficient to protect the privacy of the study participants. The next level of divisions made by the Census Bureau is called a block group. Census blocks are grouped into block groups with a nationwide average of 39 census blocks per block group. Like census blocks, block groups can vary widely in area and population. For each home within a census block that contained fewer than 10 homes, the block group was used as the geographical unit instead of the census block. None of the needed block groups contained fewer than 10 homes, so no further level of division was required.

The next step of design was how to create a map using these geographical units to display the data appropriately. The first design iteration was very simple. The map of Arizona would be drawn with each relevant geographical unit colored according to the average of the contamination data. The color scale would be a linear scale, starting with 0 ug/mg being colored in white, and the max contamination value being colored in blue. Figure 3 is an example of the first iteration of design. This example uses average concentration values for Beryllium.

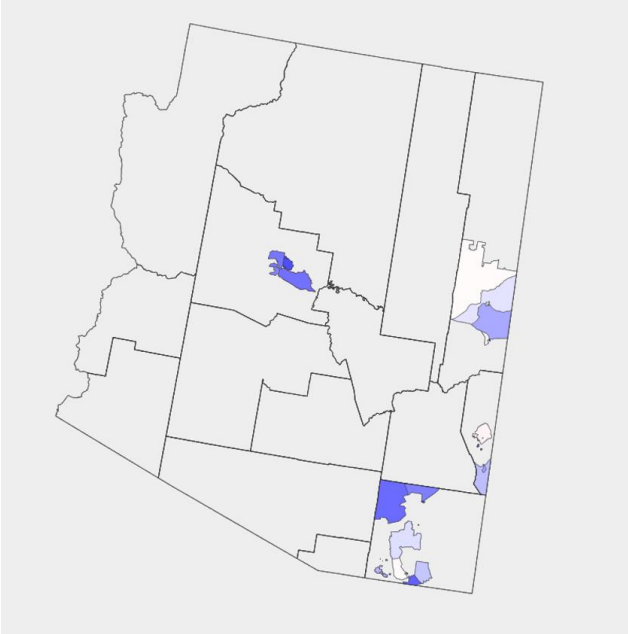

Fig 4. A screenshot of the first iteration of design of the visualization. The color scale is based on the maximum and minimum data values for Beryllium samples.

For Beryllium, the maximum average value was 3 ug/mg, so the dark blue census units show a concentration close to 3 ug/mg. The white census units show a concentration closer to 0 ug/mg. This preliminary version had many issues. First, for many of these maps, the color scale was somewhat misleading, and didn't take into account the specific characteristics of the different minerals. Second, many of the very small census blocks were invisible, while the block groups appear very large. This gives the impression that the block groups are more significant or that they contain more samples, which is not accurate. Similarly, this visualization doesn't convey the breadth of the data. This study included 398 individual collection sites, each with concentration values for 17 minerals. Last, the visualization gave no information about the distribution of values within one of those census units. The color only portrays the average value, so within a unit the values could vary widely. These were the main design issues that were addressed to produce the final visualization.

Color Scale Design

Using the max of the data as the top of the color scale meant that each time the user selected a different map, the dark blue would correspond to a different value. For example, for beryllium, the color blue corresponded to 3 ug/mg but for iron, it corresponded to 38,755 ug/mg. This may not have been a problem, as long as the changing in scale was emphasized as the user selected different mineral maps.

However, the issue became more significant with the harmful contaminants. The scale potentially gave the false impression that the concentration was at a safe or dangerous level. For example, the EPA recognizes that arsenic is potentially hazardous in soil at a concentration of 10 ug/mg. The maximum average value of this data set was 75 ug/mg.

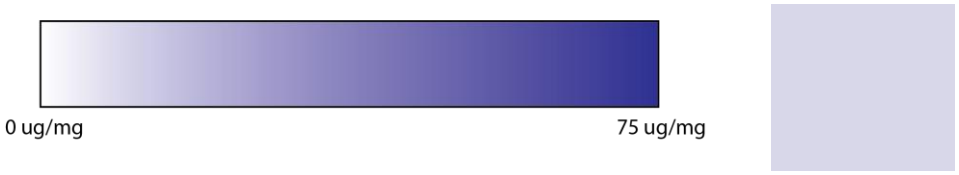

Fig 5. A sample color bar for arsenic using a linear scale based on the data.
The color on the right is the color mapping for the RILL 10 ug/mg,
Which is light blue even though it is a hazardous level.

Using the in Figure 5, the color mapped to 10 ug/mg is the colored box on the right. Any census unit with this color or darker, has an average concentration that is potentially hazardous. This is misleading because the color is so light, and users often associate saturated color with more risk or danger.

On the other hand, the EPA recognizes that barium is potentially hazardous in soil at a concentration of 15,000 ug/mg. However, the maximum average value in this data set is only 989 ug/mg. So 989 ug/mg is **not** a harmful contaminant level, but it is colored with the darkest blue. Any unit with the darkest blue color is not potentially hazardous.

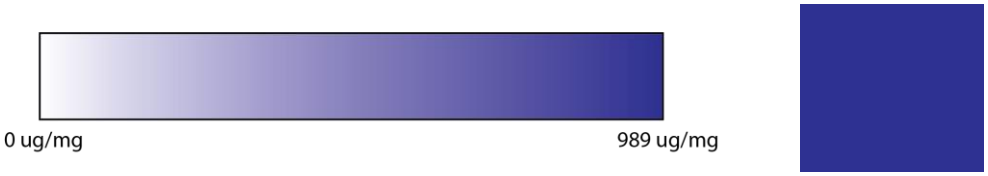

Fig 6. A sample color bar for barium using a linear scale based on the data.
The color on the right is the color mapping for the max 989 ug/mg,
which is dark blue even though 989 ug/mg is not a hazardous level.

This dichotomy in the color scales between different harmful contaminants is potentially misleading, especially if a user is switching between maps quickly and not reading into the details. Thus, this color scale needed to be redesigned to appropriately demonstrate the hazardous concentrations in a standard way. In other words, the scale needed to be consistent that dark blue always represents a hazard and white always represents zero.

The solution to this problem was to use a linear scale up from zero to the RSL, then use the same color for any value greater than or equal to that level. Also, the scale was changed to use multiple colors, because differences colors are easier to recognize than different tints of the same color. Below is the same example as above, with arsenic on the left and barium on the right. Note: major road lines were added since the last iteration to give users a sense of location.

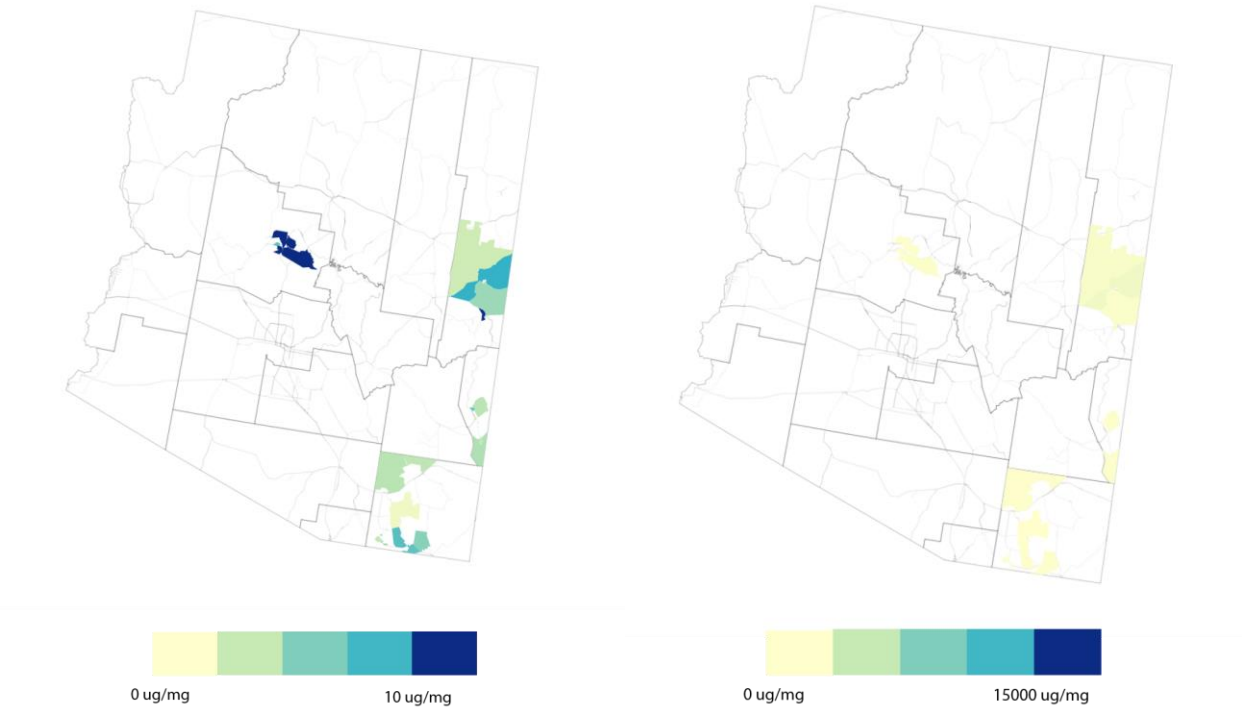

Fig 7. The redesigned color bar applied to the arsenic and barium maps.
It is clearly visible which values exceed the residential screening level.

With this scale, it is immediately clear that the arsenic values range past the potentially hazardous level, while the barium values are at very safe levels. The maps are accompanied by text explaining the meaning of the residential screening levels. For those minerals not considered hazardous, the original scale was used with a different color, so it would be obvious to the user that they type of scale is changing when they switch maps. There is accompanying text to explain that these minerals are not considered a hazard, and that the scale is a reflection of the range of the data.

Relative Size of Census Units

After the color scale was fixed, the next major issue was that certain census blocks were too small to be visible at large scale. On the other hand, the block groups were much larger and immediately visible. This conveyed that the block groups were more important or that they contained more samples than the census blocks. This was simply not true, and would be misleading to users viewing the map. A related issue was that the map did not portray the number of samples in a census block or block group. Both of these issues needed to be address by making changes to the design.

There were two initial ideas to solve this problem. The first was to draw a circle inside the boundaries of the census unit for each sample within that unit. Of course, the circles could not be drawn on their exact geo coordinates, so the circles would be evenly distributed with the boundaries of the census unit. However, we chose not to pursue this method. A few of the census blocks were simply too small, and too densely populated with collection sites for the circles to display appropriately. The same problem would arise--the circles within the block groups would be easier to see and therefore have more visual significance.

The other idea was to draw one circle representing each census unit. This circle would be colored according to the scale detailed previously. The radius of this circle would represent the number of collection sites within that unit--so census block with more sites would have more visual significance. The center of the circle would be plotted on the centroid of the census unit. For example, the following is a block group from the Barium map that is turned into a circle.

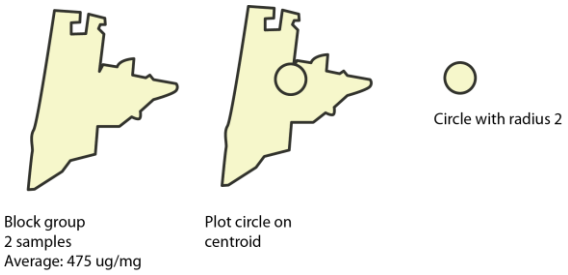

Fig 8. Shows the process of converting a census unit to a circle.

With this design, those units with the most collection sites (instead of the greatest area) have the most visual significance. The following is an example of the whole map using circles instead of census units. This method has the added advantage of further obscuring the exact location because users can't see the geographical boundaries of the census units.

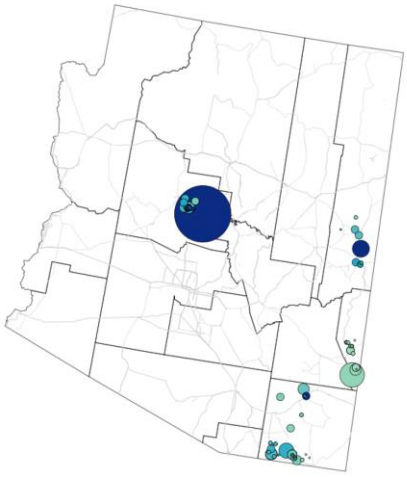

Fig 9. An example map (chromium) using the circle design.

However, using this design required adding another scale to show how the sizes of the circles corresponded to the number of samples in that unit. The following is the same map, including both the color and the circle scales.

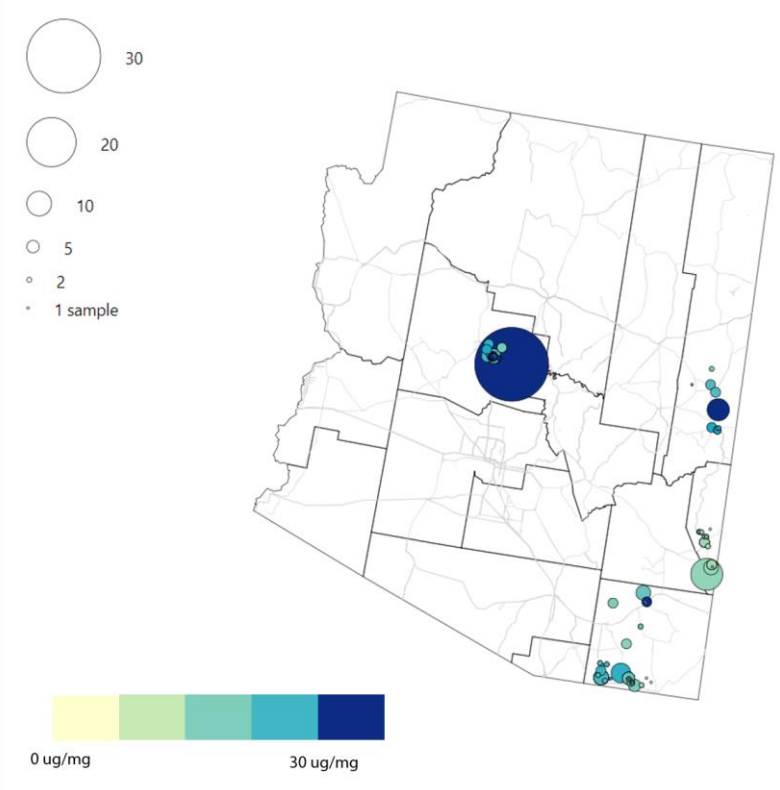

Fig 9. An example map (chromium) using the circle design, with the circle scale and color bar.

Displaying Distributions

The last design consideration was how to display the distribution within each of the circles. The initial design concept was that the map would be interactive, so the user could mouse over a circle and see a display of information about the samples in that census unit. Part of this display would inform the user the actual average value (the value used to color the circle). The other part would show the distribution. There were two different designs explored to show the distribution. The first idea was to show a histogram of all the data within that unit. This method was interesting, but it created more issues. Each histogram had different distributions of data, different bins, and different x and y axes. For example on the arsenic map, one circle had 30 samples that ranged from 0.3 ug/mg to 300 ug/mg. Another circle had 5 points all within the range of 0.1 - 0.5 ug/mg. So the axes and bin sizes changed extremely from circle to circle.

The difference between the histograms was deemed too confusing, especially since they would only be displayed on mouse over. The histograms didn't convey what was important about the distributions. Another problem with the histograms was how to color them appropriately. The bins for each histogram were different then the steps used on the color scale, so the coloring was confusing. Additionally, the initial goal for this visualization was to make the data accessible to a broad audience. Histograms are not very common outside of scientific applications. It was decided that a more simplified representation would be easier to understand and convey the appropriate message.

The second design explored was a "dot map" of the data within each circle. For this method, instead of a histogram, a number of small dots were drawn. Each dot represented one of the data points within that circle, so it was colored according to the linear scale. The advantage of this design is that it is very easy to understand, and it conveys a unified message across the different circles. The disadvantage was that for the circles with more unified distribution, the dot map was not very interesting compared to the histogram. Between the two options, it was decided that the dot map conveyed the distribution information better than the histograms. The following figure shows two examples of this dot map distribution.

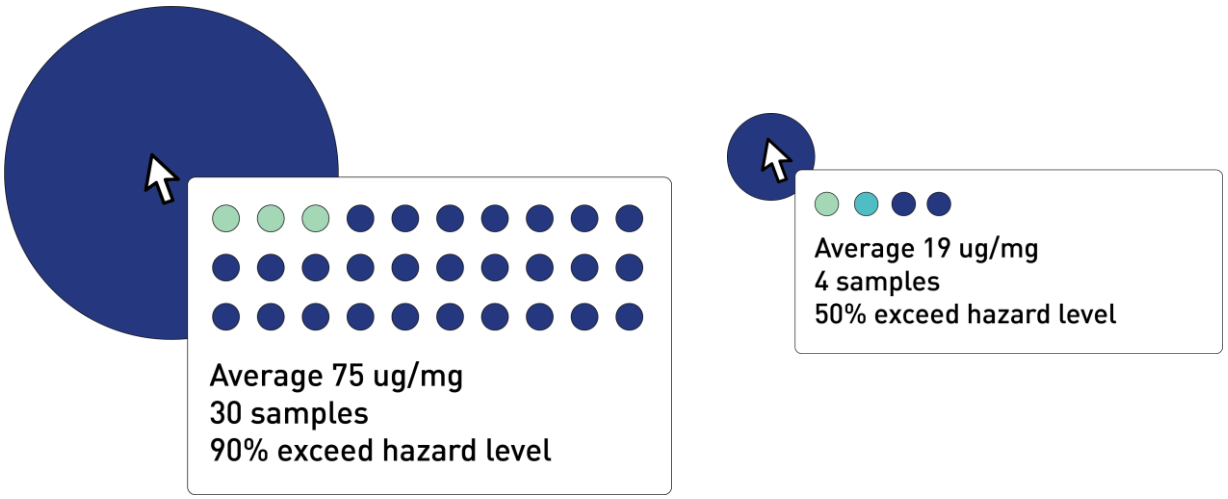

Fig 10. Examples of the distribution tooltip, using two circles from the arsenic map

.

To summarize, this visualization was designed starting with geographical units set by the US Census Bureau. For each collection site, the census block was used to obscure the exact location. For those blocks with fewer than 10 households, the block group was used instead. These areas were colored according to one of two different scales. If the map was for a mineral considered hazardous, a linear color scale between zero and the residential screening level was applied to the average to produce a color. Any values exceeding the RSL were colored with the maximum navy blue color. This was done to convey the same information about hazard levels

across the maps. If the map was for a mineral not considered hazardous, it was colored according to a simpler linear scale between zero and the maximum of the data. A different color was used to show clearly when the scales change between maps. To accurately convey the breadth of the study and the number of samples within a census unit, the census units were transformed into circles with a radius of the number of samples. These circles were plotted at the centroid of the census unit. Last, to display the distribution within one such circle, a tooltip was designed with an array of dots. Each dot represents a single sample, colored according to the same scale as the larger circles. Overall, the main goal of this design was to convey the data in a simple, aesthetically pleasing, and visually engaging way, while still protecting the privacy of the original study participants

Methodology

To build this visualization I used a combination of web tools and command line tools. The soil contamination data from the study was in excel sheets, including the participants' physical address and the concentration values for each mineral. The first step of the data processing was to convert the physical addresses to geo coordinates so they could be plotted on a map later on. This was done using a free service that converts street addresses to latitude and longitude coordinates. Next, the coordinates were entered in the excel file alongside the concentration data, and converted into JSON. At this point, the data consisted of a JSON array of objects; each object included coordinates, county name, and the concentration data.

The next step was to acquire the shapefile map data from the US Census Bureau. First a simple map of Arizona with the county divisions was obtained. Then, a very large shapefile including the divisions of every single Arizona census block was obtained. This was immediately filtered (to reduce the file size) to only include the counties of interest: Apache, Cochise, Greenlee, and Yavapai. Last, a similar shapefile with just the block groups was obtained. The files for the blocks and block groups also contained the population and housing information appended to each block/group. Each of these shapefiles was converted to JSON using Nodejs. The following image is the census block shapefile with only the four counties of interest.

Fig 11. Map of all census blocks in Apache, Cochise, Yavapai, and Greenlee Counties.

Next the census blocks that contained one or more collection site were isolated from this map. This was done using Nodejs.

Fig 12. Map of all census blocks that contained one or more collection sites.

At this point, each census block was examined to see it had fewer than 10 households. Each collection site that existed inside a census block with fewer than 10 households was isolated. These sites were then located within their block group instead. The list of census blocks and block groups was then merged, and merged with the Arizona county map to produce the following map.

Fig 13. Map of all relevant census blocks, and block groups merged with the Arizona counties, used for the first design iteration.

This is the map that was used for the first iteration of design. Each census unit (block or group) had a list of the collection sites within that unit, and their corresponding concentration values. The averages were calculated on each census unit, for each mineral. To generate the circles in the final design, the centroid of each census unit was calculated using Nodejs. The map was then plotted on the website using D3.js version 3. The

other elements of the final design, including the distribution tooltip, the changing scale, and the map selector, were also completed using D3.js and Javascript.

Next Steps

The next step for this project is to design a survey to engage the study participants with the visualization. The main goals of the survey would be to ensure that A) the privacy of the participants is protected appropriately and B) the data is displayed in a simple and interesting way that any user can understand. After the survey is completed and feedback is collected from the participants, design changes will likely be made to the visualization in response to that feedback. After these changes, the final design will also be approved by the IEHSRL team. Then the same design and methodology will be applied to the water data set, including minor changes to the design to emphasize the difference between the two data sets.

Before deploying these visualizations, some other functionality will be added to the web interface. There will include an address lookup, so web users can see where they live in relationship to the collection sites. This lookup will also indicate whether their home falls within one of the relevant census units. This added functionality will encourage users to think about how this data is applicable to their lives and homes, and will hopefully encourage people to get involved with citizen science. Finally, the two visualizations will deploy to the Gardenroots website.

Related Work

Before this thesis, the IEHSRL team created the Gardenroots Citizen Science Project, including initial community partnership, training study participants, and final report-back of data to the community. Details on the creation and execution of the Gardenroots study, and the subsequent communication can be found in Ramirez-Andreotta et al. The laboratory designed and distributed booklets for each individual participant, giving a detailed analysis of that participant's data in the context of the study. The booklets can be found at gardenroots.arizona.edu.

Beyond print media, technology and data visualization provides a unique opportunity for the field of citizen science to engage study participants and the general public. There is a growing body of work similar to Gardenroots that utilizes data visualization for the purpose of reporting back to citizen participants. Hochachka et al investigates the use of data science for mapping ecological data on a large scale. The study focuses specifically on ensuring data quality is consistent and data quantity is maximized. Newman et al presents a general outline for citizen science projects including the use of emerging technologies including data visualization, mobile apps, and web interfaces.

In the field of data visualization, there is limited research on privacy-preserving visualization. Many of the privacy related works focus more on data mining, and how to protect privacy during data collection and storage, rather than during visualization. However there are a few papers that have a similar focus as this thesis, examining how to use data science and population information to obscure exact locations when designing mapping visualizations. Research from Okansen et al, published in the Journal of Transport Geography, explores generating heat maps from mobile sports tracking data. This study focuses specifically on how to limit bias caused by varying availability of data in certain areas of the map. Machanavajjhala et al provides a different approach, generating a map of US commuting patterns using synthesized data that is statistically similar to the original data. Last, Chen et al presents a broad method of generating a privacy preserving map for any type of crowd-sourced data, by requiring specific user inputs. The user uploads sparse, unorganized locations, which the server then uses to create the map instead of the user's exact location. In general, each approach attempts to generalize the data, while making the visualization still communicate the same message that it did with the real data.

Ramirez-Andreotta, M. D., Brusseau, M. L., Artiola, J., Maier, R. M., & Gandolfi, A. J. (2015). Building a co-created citizen science program with gardeners neighboring a superfund site: The Gardenroots case study. International Public Health Journal, 7(1), 13.

Ramirez-Andreotta, M. (2012). Designing a comprehensive, integrated approach for environmental research translation: The gardenroots project to empower communities neighboring contamination (Order No. 3548730). Available from Dissertations & Theses @ University of Arizona; ProQuest Dissertations & Theses Global: Science & Technology; ProQuest Dissertations & Theses Global: Social Sciences. (1271756486).

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012). The future of citizen science: Emerging technologies and shifting paradigms. Frontiers in Ecology, 10(6).

Wesley M. Hochachka, Daniel Fink, Rebecca A. Hutchinson, Daniel Sheldon, Weng-Keen Wong, Steve Kelling, Data-intensive science applied to broad-scale citizen science, Trends in Ecology & Evolution, Volume 27, Issue 2, 2012, Pages 130-137, ISSN 0169-5347, https://doi.org/10.1016/j.tree.2011.11.006.

Juha Oksanen, Cecilia Bergman, Jani Sainio, Jan Westerholm, Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data, Journal of Transport Geography, Volume 48, 2015, Pages 135-144, ISSN 0966-6923, https://doi.org/10.1016/j.jtrangeo.2015.09.001.

X. Chen, X. Wu, X.-Y. Li, Y. He, Y. Liu, "Privacy-preserving high-quality map generation with participatory sensing", Proceedings of IEEE INFOCOM, 2014.

Ashwin Machanavajjhala , Daniel Kifer , John Abowd , Johannes Gehrke , Lars Vilhuber, Privacy: Theory meets Practice on the Map, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, p.277-286, April 07-12, 2008

Appendix

To produce the visualization, the following data processing steps were performed.

1. Collect the data into a single CSV file linking addresses to concentration data.
2. Use address lookup to convert addresses into latitude/longitude pairs (LatLong.net).
3. Convert CSV file into JSON (convertcsv.com).
4. Obtain census block data, block group data, and county data from US Census Bureau (https://www.census.gov/geo/maps-data/data/tiger.html). These should be zipped files including a .shp file.
5. Convert cb data, bg data, and county data .shp files into json using shp2json. At this point each file will be a geojson file, including a bounding box, some properties, and an array of features.
6. Cut the cb json file to just a feature array using head, tail, and cut. Keep the prefix and suffix in separate files to remerge later. This step is necessary because the file is too large to be read via javascript file reading, and must be read through an array stream.
7. Isolate the census blocks in the four relevant counties. This is done to reduce the file size and thus make each subsequent task take less time. This must be done using stream-json utils to read the file.
8. Remerge the data with the cut prefix and suffix. At this point, the file should be valid geojson including the census blocks for each of the four relevant counties. This can be validated by uploading the geojson file at mapshaper.org.
9. Isolate the census blocks that contain one or more collection sites. To do this, read in the json file with the lat/long and concentration. Use geojson-utils point-in-polygon to locate the proper census block for each data point. Isolate these census blocks and write them back to geojson, with a list of points appended to each census block. Exclude any census blocks with no data points.

10. Identify those census blocks that have fewer than 10 households. For each of these subpar census blocks, removed them from the dataset, and collect the data points contained in them. At this point, there is a file with only the well-populated census blocks (and appended data points), and a separate file with a list of data points from the unpopulated census blocks.
11. Repeat step 9 but instead isolate the **block groups** that contain one or more of the data points from the unpopulated census blocks. This should generate a valid geojson with a list of block groups, with a list of data points appended to each block group.
12. Merge the two files (block group file and census block file) using geojson-merge. At this point, the file should be valid json and should be sufficient to produce the first design iteration detailed in the design section of this thesis.
13. Use geojson-polygon-center to find the centroid of each census unit (block or group) and write that to a separate file including the concentration data.
14. Iterate over the centroid data and calculate the averages of each concentration value, for each mineral, for each centroid. Also count the number of points and append that to the centroid object.
15. Finally, plot the original county map in D3.js, and overlay it with a map of circles from the centroid data. Use the count for the radius of the circle, and the average value for the fill color.